Dual-path Image Inpainting with Auxiliary GAN Inversion

Wentao Wang[1], Li Niu[1*], Jianfu Zhang[2], Xue Yang[1], Liqing Zhang[1*]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2] Tensor Learning Team, RIKEN AIP
{wwt117,ustcnewly,yangxue-2019-sjtu,lqzhang}@sjtu.edu.cn, jianfu.zhang@riken.jp

The supplementary material contains the following parts: (1) Additional Qualitative Comparison: provide more visual comparison results on FFHQ [1], LSUN cat, and LSUN church [7] with regular and irregular holes; (2) Additional Quantitative Comparison: provide additional quantitative comparisons on LSUN church; (3) Model Complexity and Inference Time; (4) User Study; (5) Additional Experiments on Places2; (6) Additional Ablation Study.

1. Additional Quantitative Comparison

We conduct additional quantitative comparison with Yeh *et al*. [5], Lahiri *et al*. [3], PICNet [10], GC [8] on LSUN church. The test setting on LSUN church is the same as that on FFHQ. Four evaluation metrics: relative l_1, Structural Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Frechet Inception Distance (FID) are adopted. The evaluation results on LSUN church are shown in Table 1. It can be seen that our method also outperforms other methods for all evaluation metrics and all mask ratios.

	Mask	Yeh *et al*. [5]	Lahiri *et al*. [3]	GC [8]	PICNet [10]	Ours
ℓ_1 (%)↓	0-10%	0.69	0.78	0.56	0.60	**0.54**
	10-20%	1.47	1.82	1.36	1.38	**1.24**
	20-30%	2.89	3.23	2.49	2.43	**2.22**
	30-40%	4.11	4.73	3.75	3.58	**3.24**
	40-50%	5.75	6.44	5.21	4.97	**4.41**
	50-60%	8.03	8.99	7.28	7.24	**6.11**
	Ave%	3.82	4.33	3.44	3.37	**2.96**
SSIM↑	0-10%	0.962	0.960	0.967	0.965	**0.969**
	10-20%	0.907	0.899	0.918	0.912	**0.921**
	20-30%	0.835	0.823	0.850	0.846	**0.858**
	30-40%	0.761	0.747	0.779	0.778	**0.793**
	40-50%	0.695	0.668	0.701	0.702	**0.723**
	50-60%	0.593	0.587	0.614	0.607	**0.643**
	Ave%	0.792	0.781	0.805	0.802	**0.818**
PSNR↑	0-10%	30.958	29.851	31.894	31.711	**32.578**
	10-20%	25.540	24.758	26.215	26.493	**27.235**
	20-30%	22.167	21.717	22.899	23.480	**24.148**
	30-40%	19.975	19.616	20.663	21.412	**22.071**
	40-50%	18.443	17.846	18.900	19.671	**20.413**
	50-60%	16.736	15.789	17.005	17.523	**18.567**
	Ave%	22.303	21.596	22.929	23.383	**24.169**
FID↓	0-10%	1.11	1.42	0.84	0.98	**0.77**
	10-20%	3.83	4.79	2.59	3.02	**2.30**
	20-30%	9.03	11.58	5.73	6.76	**4.92**
	30-40%	16.44	19.85	10.39	11.99	**8.28**
	40-50%	25.23	29.63	17.19	19.35	**12.95**
	50-60%	27.88	32.83	27.24	29.49	**17.81**
	Ave%	13.92	16.68	10.66	11.93	**7.84**

Table 1. Quantitative comparison on LSUN church [7].

2. Additional Qualitative Comparison

We conduct qualitative comparison with Yeh *et al*. [5], Lahiri *et al*. [3], PICNet [10], GC [8], CoModGAN [9] on FFHQ, LSUN church, and LSUN cat with regular and irregular holes. The comparisons are shown in Figure 5, Figure 6 and Figure 7. Since the face image are already aligned in FFHQ dataset, it is easy for all methods to learn the valid information over FFHQ dataset and to generate relatively reasonable results. In LSUN cat and LSUN church dataset, the images are diverse and complex. Compared to feed-forward inpainting, GAN inversion inpainting and our method can generate results with better semantics. For example, the feed-forward inpainting methods like PICNet and GC often fail to generate cat faces (*e.g*., row 1, 4 in Figure 7), which requires necessary semantic knowledge. In contrast, in our method, the inversion path can provide useful semantic knowledge for the feed-forward path to generate well-structured and semantically meaningful cat faces. GAN inversion inpainting methods are prone to generate results having color discrepancies without post-processing. Our method outperforms other methods on three test sets with the most reasonable and realistic results.

3. Model Complexity and Inference Time

We compare our model complexity and inference speed with other baseline methods in Table 3. In Yeh *et al*. [5], we optimize 1,000 times to generate the final results. The model size of StyleGAN2 [2] pretrained model used in Yeh *et al*. [5], Lahiri *et al*. [3] and ours is 30.03M. We denote it as "S" in Table 3 for simplification.

It can be seen that Yeh *et al*. needs the longest infer-

*Corresponding author.

	ℓ_1 (%)$^{\downarrow}$				SSIM$^{\uparrow}$				PSNR$^{\uparrow}$				FID$^{\downarrow}$			
Mask(%)	0-20	20-40	40-60	Ave	0-20	20-40	40-60	Ave	0-20	20-40	40-60	Ave	0-20	20-40	40-60	Ave
HiFill	2.90	5.40	9.32	5.87	27.041	21.653	17.872	22.188	0.889	0.751	0.584	0.741	3.58	16.95	55.44	25.32
MEDFE	1.95	3.26	6.61	3.94	29.564	24.563	19.975	24.701	0.942	0.851	0.698	0.830	2.36	10.79	34.84	16.00
CoModGAN	1.75	3.70	8.03	4.49	29.906	22.599	18.507	23.671	0.945	0.824	0.620	0.796	2.11	6.58	15.85	8.18
Ours	**1.58**	**3.23**	**6.44**	**3.75**	**30.033**	**24.587**	**20.223**	**24.948**	**0.949**	**0.857**	**0.705**	**0.837**	**1.83**	**5.65**	**14.22**	**7.24**

Table 2. Quantitative comparison on Places2 [11].

	Yeh *et al.*	Lahiri *et al.*	GC	PICNet	CoModGAN	Ours
Speed (s/frame)	45	0.088	0.024	0.080	-	0.093
Param. (M)	S + 0	S + 4.5	10.0	6.04	109	S + 28.05

Table 3. Comparison of model complexity and inference speed. "S" is 30.03M, which denotes the model size of pretrained StyleGAN2.

ence time due to multiple optimizations of the latent code. Our method has comparable inference speed with Lahiri *et al*. [3] and PICNet [10]. Since CoModGAN is tested with image size of 512, we omit its inference speed. Although the performance of CoModGAN is similar to ours, it has the largest model size. We acknowledge that the model size of our method is larger than other three methods because of dual-path architecture, but our method can generate more realistic and reasonable results.

4. User Study

Following [8], we conduct user study on 200 images randomly selected from two datasets, in which each image is processed with regular or irregular masks. 30 subjects with basic background in computer vision are invited to rank the subjective visual qualities of images. We perform three pairwise comparisons for each baseline with our method: (1) Our method *v.s.* Yeh *et al*., (2) Our method *v.s.* Lahiri *et al*., (3) Our method *v.s.* PICNet, (4) Our method *v.s.* GC. We omit CoModGAN because we can not obtain its results on LSUN dataset. A total of $200 \times 30 = 6,000$ comparisons were conducted for each baseline. The study shows that 80.10% (4,806 out of 6,000), 85.32% (5,119 out of 6,000), 82.17% (4,930 out of 6,000), and 82.58% (4,955 out of 6,000) of comparisons preferred our results over Yeh *et al*., Lahiri *et al*., PICNet, and GC, respectively.

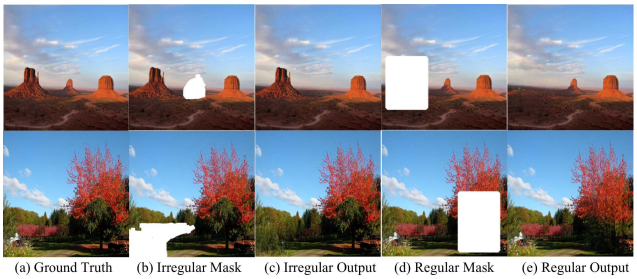
5. Additional Experiments on Places2

We have proved the effectiveness of our method on relatively homogeneous and aligned datasets (*e.g*., FFHQ, LSUN) which have complex semantic. However, it is also necessary to verify our method on diverse domains dataset like Places2 [11]. Thus, we compare our method with three methods: HiFill [6], MEDFE [4], CoModGAN [9] on Places2 . We first train StyleGAN2 model from scratch on Places2 for about two weeks and Figure 1 show some results randomly generated by StyleGAN2 model trained on Places2.

Figure 1. Some results randomly generated by StyleGAN2 model trained from scratch on Places2.

Table 2 shows our method quantitatively compares with three methods. The test setting is the same as the FFHQ in the main text while 10,000 images are randomly selected for testing. It can be seen that our method is superior to other methods among all metrics.

We also provide some examples for real inpainting application based on our method in Figure 2. In the experiment, we observe that the inversion path can generally provide useful semantic prior for inpainting on relatively diverse dataset like Places2.



(a) Ground Truth (b) Irregular Mask (c) Irregular Output (d) Regular Mask (e) Regular Output

Figure 2. Some examples for real inpainting application based on our method.

6. Additional Ablation Study

In this section, we investigate the effect of the number of our proposed deformable fusion module layers on the final results. We also provide more experiments to prove the effect of the inversion path on the final results and the analyses of deformable fusion module.

Settings	ℓ_1 (%)$^{\downarrow}$	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	FID$^{\downarrow}$
6-DF	2.16	0.883	28.087	4.05
2-CF and 4-DF	2.17	0.883	28.081	4.07
3-CF and 3-DF (Ours)	2.17	0.882	28.078	4.02
4-CF and 2-DF	2.20	0.879	27.884	4.26
5-CF and 1-DF	2.23	0.856	27.823	4.55
6-CF	2.29	0.870	27.818	4.73

Table 4. Additional ablation studies for the number of deformable fusion module layers. The setting (c) is our method.

6.1. Additional Experiments on the Number of Deformable Fusion Module Layers

Since there are six layers in our generator, we let the layer with the lowest resolution (8×8) be the first layer. We compare with the six following settings: (a) All layers use deformable fusion ("6-DF"); (b) the first two layers use concatenation fusion and the last four layers use deformable fusion ("2-CF and 4-DF"); (c) the first three layers use concatenation fusion and the last three layers use deformable fusion ("3-CF and 3-DF"); (d) the first four layers use concatenation fusion and the last two layers use deformable fusion module layers ("4-CF and 2-DF"); (e) the first five layers use concatenation fusion and the last layer use deformable fusion ("5-CF and 1-DF"); (f) All layers use concatenation fusion ("6-CF"). All the results are tested on FFHQ datasets and the experimental setup is the same as Section 4.2 in the main text. Note that setting (c) is actually our method. The results are summarized in Table 4. Compared with simple concatenation fusion (setting (f)), our proposed deformable fusion module promotes the final results. Specially, when the deformable fusion module is applied to high-resolution (256×256, 128×128, 64×64), the promotions are apparent, especially on FID (setting (c) (d) (e)). However, when the deformable fusion module is applied to lower-resolution (32×32, 16×16, 8×8), the promotions are limited (setting (a) (b)). Taking accuracy and efficiency into account, we choose setting (c), which has the lowest FID, as our final method.

6.2. Additional Experiments on Dual Path

We provide more experiments to prove the effectiveness of the inversion path on the final results. Figure 3 shows the visualization comparison between the inversion path and the feed-forward path. The inversion path network and the feed-forward path network are trained separately. The results of inversion path and dual-path in Figure 3 is the outputs generated by the inversion path of our method and the outputs generated by the feed-forward path of our method. It can be seen the inversion path can provide extra semantic information (*e.g.*, cat face and eyeglasses) for the feed-forward path and have a positive influence on the final results. Without the assistance of the inversion path, there are high chances for single feed-forward path network to generate poor results. (*e.g.*, Single feed-forward path network can not restore the cat face.) These results prove that the inversion path is able to provide extra semantic prior and improve the inpainting results.

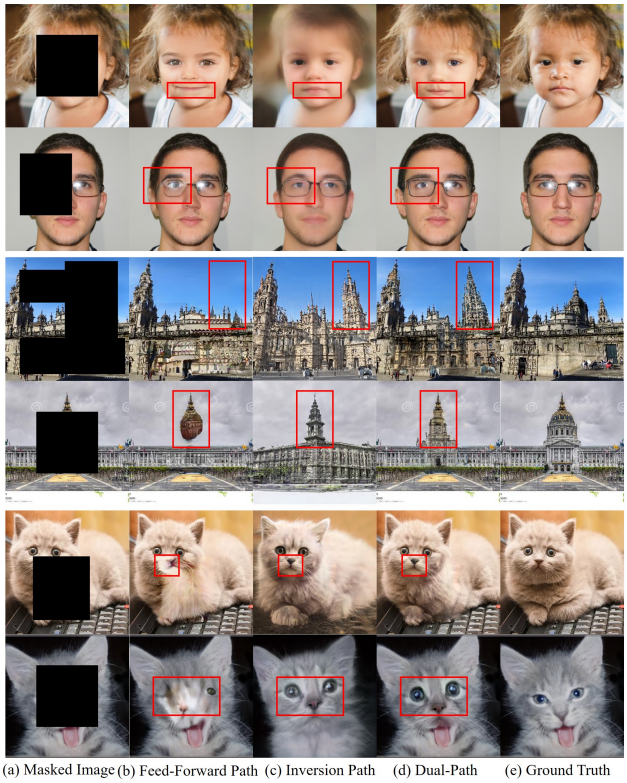
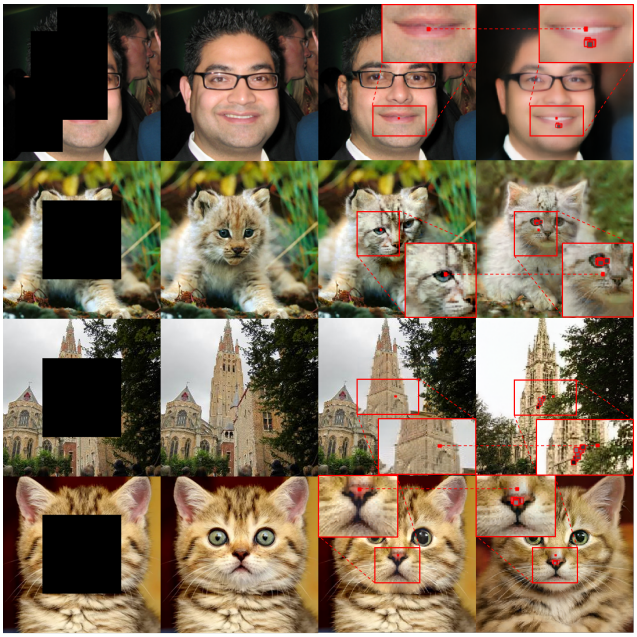


Figure 3. Visualization of Two Path Output. The differences are highlighted in red boxes. Comparing (b), (c) and (d), we can find the inversion path provide semantic prior (*e.g.*, cat face and eyeglasses) assisting for the feed-forward path. Best viewed by zooming in.

6.3. Analyses of Deformable Fusion Module

In Figure 4, we provide four examples of FFHQ, LSUN church and LSUN car to verify the effectiveness of deformable fusion module. Similar to Sec 3.2 in the main text, we visualize the outputs at the largest resolution (256×256) from two paths. For each example, we select one point in the feed-forward path as the target point (red point). We also draw the target point (red point) at the same location in the inversion path. By comparing the target points in two paths, we can clearly see the misalignment issue between two paths.

Recall that we use 3×3 deformable convolution kernel with learnable offsets and modulations. We visualize the learnt offsets and modulations of the kernel centering at the target point in the inversion path. Specifically, we draw the 9 deformed sampling points with the corresponding modu-



(a) Masked Image (b) Ground Truth (c) Feed-Forward Path (d) Inversion Path

Figure 4. Visualization of the effectiveness of deformable fusion module. The deformable fusion module can attend to the correct information (*e.g.*, lips in row 1, cat eyes in row 2, buildings in row 3 and cat noses in row 4) in the inversion path. Best viewed by zooming in.

lations (marked with grey values). It can be seen that our deformable fusion module can attend to the relevant information in the inversion path to avoid the misalignment issue. For example, in row 1, it attends to the lip instead of the tooth. In row 2, it attends to the cat eye instead of the cat face. In row 3, it attends to the building instead of the tree. In row 4, it attends to the cat nose instead of the cat face. These results demonstrate that our deformable fusion module can alleviate the misalignment issue and help the feed-forward path incorporate more compatible information from the inversion path.

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 5

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1

[3] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13696–13705, 2020. 1, 2

[4] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, pages 725–741, 2020. 2

[5] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 1

[6] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 2

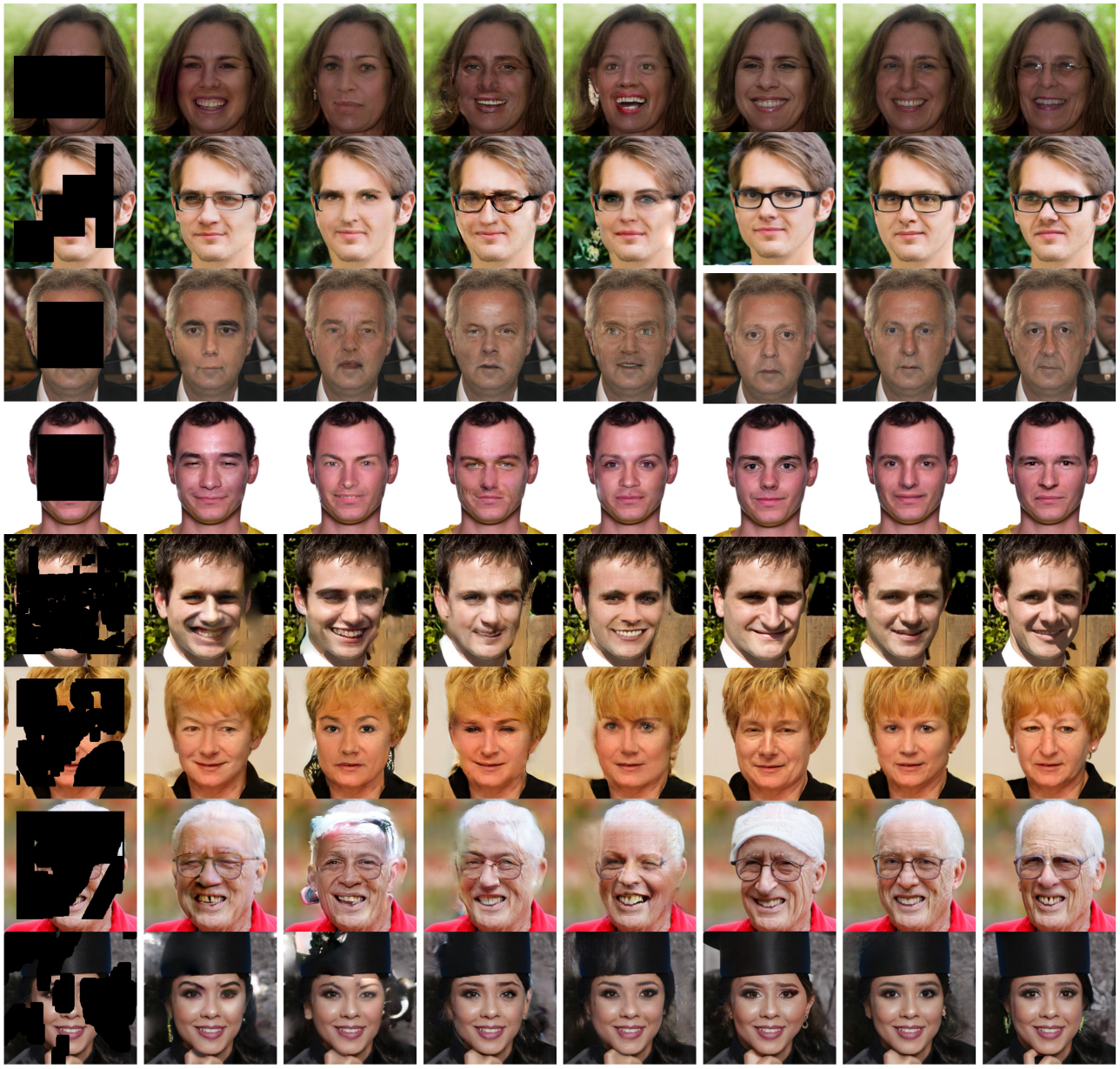
[7] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 6, 7

[8] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2

[9] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 1, 2

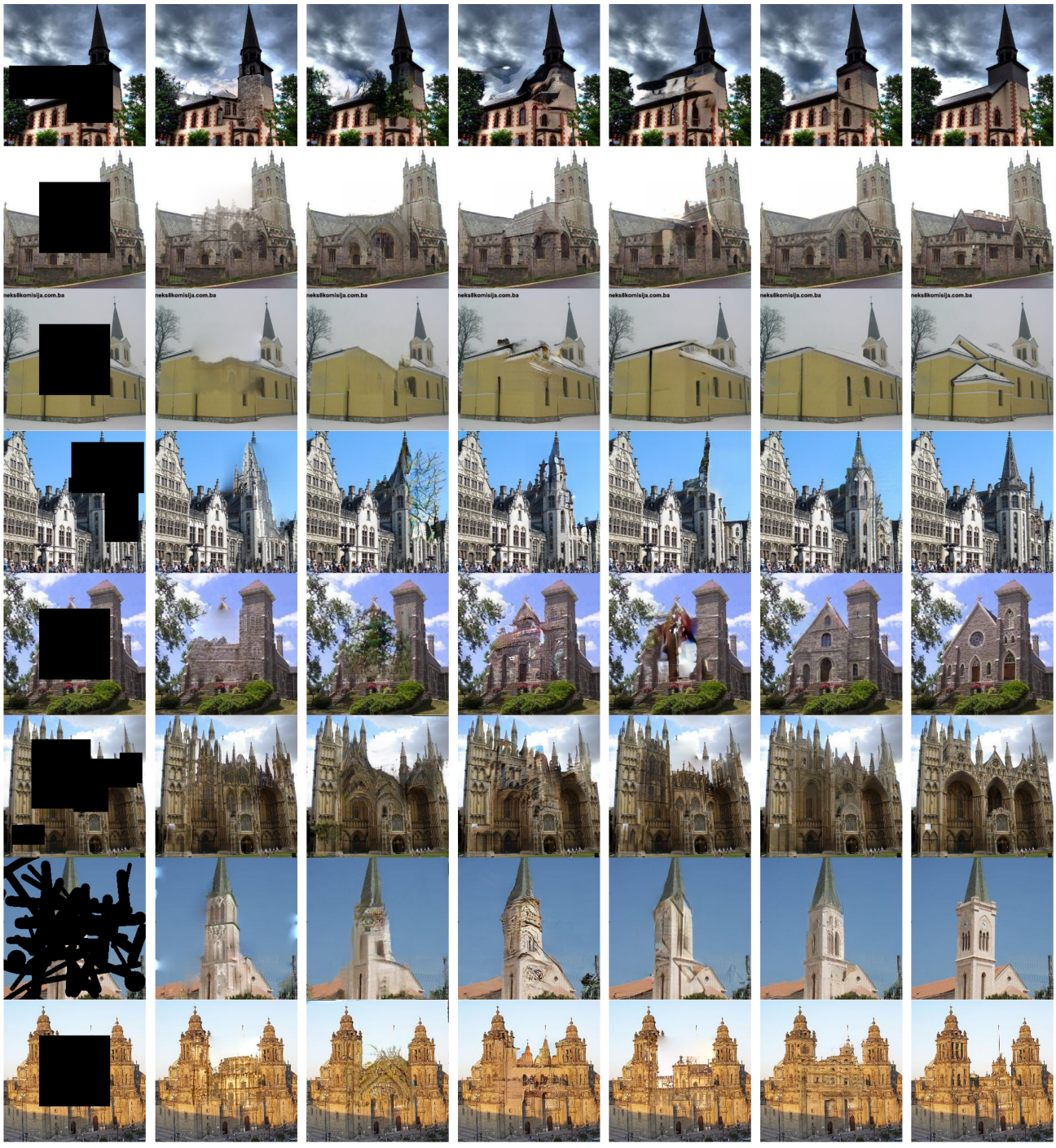
[10] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 1, 2

[11] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2



(a) Masked Image (b) Yeh et al. (c) PICNet (d) GC (e) Lahiri et al. (f) CoModGAN (g) Ours (h) Ground Truth

Figure 5. More visual comparison results on FFHQ [1].



(a) Masked Image (b) Yeh et al. (c) PICNet (d) GC (e) Lahiri et al. (f) Ours (g) Ground Truth

Figure 6. More visual comparison results on LSUN church [7]

(a) Masked Image (b) Yeh et al. (c) PICNet (d) GC (e) Lahiri et al. (f) Ours (g) Ground Truth

Figure 7. More visual comparison results on LSUN cat [7].